SPAM EMAIL DETECTION USING HYBRID FEATURES

[1]K.Arun Kumar, [2]Y.Amrutha, [3]K.Sreevani,
[4]K.Ram Kumar, [5]M.Harish,
[1,2,3]Assistant Professor, Dept. of CSE, [4,5] B. Tech., (CSE)
Malla Reddy Engineering College (Autonomous), Secunderabad, Telangana State

Abstract

As an attack of social engineering, phishing email has caused tremendous financial loss to recipients. Therefore, there is an urgent need for phishing email detection with high accuracy. In this project, we propose phishing emails detection based on hybrid features. By analyzing the email-header structure, email-URL information, email-script function and email psychological features, we extract hybrid features. Then we choose Support Vector Machine (SVM), LSTM and CNN classifiers to evaluate our experiments. Experiments are performed on a dataset consisting of legitimate emails and phishing emails. The proposed approach achieves overall true-positive rate, false-positive rate, precision and accuracy. The results show that psychological features can improve the accuracy of detection and reduce the false-positive rate. Our proposed method has a good performance in detecting phishing emails.

Keywords: -Phishing,SVM, LSTM,CNN

1. INTRODUCTION

Electronic-mail (Email) is one of the most effective and effortless sources to convey messages. Email is regarded as the cautious communication like sending of messages over networks and is an inexpensive method. Even though there are many modes to convey messages, email demand didn't reduce mainly in business, education sectors and other private and government sectors as email is treated as protected manner of conveying message. Email services playa major part in everyone's life. Compared to olden days, nowadays everyone is using email effectively. There is a gradual increase in users compared to 2015 in 2016. Nearly 5.6 billion people are operating email in 2017 and observations shows that the number will rise to 6 billion users in coming years over other apps like [RH11]. Major problem with email has been spamming and phishing mails which causes data theft by some virus attacks and are used in trickery schemes, advertisements etc. By observing previous years, the inappropriate emails, email phishing and spamming has raised recently and many privacy data threats evolved and cause huge damages to profession, each person and finance. Particularly for profession business emails inspect and examine these transmission networks can reveal secrets and private data with hard patterns of processes and decision making. Detecting these spam/Inappropriate Emails precisely in transmission networks is essential. Phishing mails are type of spam mail which are dangerous to users. A phishing mail can theft our data without our understanding once it is accessed. Thus, knowing spam mails from phishing mails is necessary. One way to protect our data from inappropriate spam mail is to create a strong password with characters, numbers and special characters, and also having a secondary password to login credentials. Can also try another manner is to alarm the user once a Spam mail tries to theft the user's data.

2. RELATED WORK

In general, there are two methods for detecting phishing emails: the blacklist technique and the machine-learning method. The sender blacklist and the URL blacklist are both part of the blacklist feature library. If several blacklist words appear in the email's subject or text, it is likely to be a phishing email. Despite the simplicity and efficacy of the blacklist, it was unable to detect fresh phishing assaults. Meanwhile, collecting blacklists takes time. The system uses machine learning to classify fresh phishing emails. Among the available phishing email detection algorithms, these have the highest detection precision and efficiency.

We proposed email detection using Support Vector Method (SVM), Long short term memory (LSTM) and Convolution Neural Network (CNN) algorithms. As these three ML algorithms are used for classification purpose, these algorithms are used for getting high accuracy.Long Short Term Memory Network is an advanced RNN, a sequential network that allows information to persist. It is capable of handling the vanishing gradient problem faced by RNN. A recurrent neural network is also known as RNN is used for persistent memory.Let's say while watching a video you remember the previous scene or while reading a book you know what happened in the earlier chapter. Similarly RNNs work, they remember the previous information and use it for processing the current input. The shortcoming of RNN is, they cannot remember Long term dependencies due to vanishing gradient. LSTMs are explicitly designed to avoid long-term dependency problems.

A convolutionneural network, or CNN, is a deep learning neural network sketched for processing structured arrays of data such as portrayals.CNN are very satisfactory at picking up on design in the input image, such as lines, gradients, circles, or even eyes and faces.This characteristic that makes convolutional neural network so robust for computer vision.CNN can run directly on a underdone image and do not need any pre-processing.A convolutional neural network is a feed forward neural network, seldom with up to 20.The strength of a convolutional neural network comes from a particular kind of layer called the convolutional layer.CNN contains many convolutional layers assembled on top of each other, each one competent of recognizing more sophisticated shapes.With three or four convolutional layers it is viable to recognize handwritten digits and with 25 layers it is possible to differentiate human faces.

3. IMPLEMENTATION

Data Collection

Collecting email data, which includes both valid and phishing emails; this data can be collected using email codes, or datasets can be accessed on Google; these data sets can be used for further application, such as extracting email data, which can then be used to train email data.

Annotation of Data

The classification of the dataset into valid and spam emails must be done manually, according to data annotation. Large amounts of training data are required to create an AI or machine learning model that behaves like a human. A model must be trained to grasp specific information in order to make judgments and take action.

ThecategorisingandlabellingofdataforAIapplicationsisknownasdataannotation.Foraspecificusecase, trainingdatamustbecorrectlyclassifiedandannotated. Companiesmay establish handimproveAIsolutionsbyusinghigh-quality,human-powereddata annotation.Asaresult,customerswillbenefitfromfeaturessuchasproductsuggestions,relevantsearchengineresults,machine vision,speechrecognition,Chabot's,andmore. Text,audio,image,andvideoaretheforuminformsofdata.

TextAnnotationTextisthemostoftenseddatatype,with70percentofenterprises relyingo nit,accordingtothe2020StateofAIandMachineLearningreport.Textannotationsinclude sentiment,intent,andquestion,amongthethings.WehaveSentimentAnnexationisttext:Sentimentanalysisevaluatesattitudes,emotions,and opinions,thereforehavingthecorrecttrainingdiacritical.Humaninnotatorsarefrequentlyusedtoacquirethatdata'sinCatheycannilysentiment andfiltercontentionallwebplatforms,includingsocialandecommerce sites,aswellastagandreportonprofane,sensitive,oncologicphrases.

Audio Annotation

The transcription and time-stamping of voice data, including the transcription of precise pronunciation and intonation, as well as the identification of language, dialect, and speaker demographics, is known as audio annotation. Every use case is unique, and some require a highly particular methodology, such as the tagging of aggressive speech indicators and non-speech sounds such as glass breaking for use in security and emergency hotline technologies.

Video Annotation

Human-annotated data is essential for successful machine learning. Humans are simply better than machines at managing subjectivity, understanding purpose, and coping with ambiguity. When determining whether a search engine result is relevant, for example, a large number of people must be involved in order to reach a consensus. When training a computer vision or pattern recognition system, humans are required to recognize and annotate specific data, such as emphasizing all pixels in a picture that contain trees or traffic signs. Using this structured data, machines can learn to spot these relationships in testing and production.

Image Annotation

Image annotation is critical for a variety of applications, including computer vision, robotic vision, facial recognition, and machine learning-based image interpretation. Metadata in the form of identifiers, captions, or keywords must be supplied to the photos in order to train these solutions. There are many use cases that demand large numbers of annotated photos, ranging from computer vision systems used by self-driving vehicles and machines that pick and sort product to healthcare applications that auto-identify medical disorders. By efficiently training these systems, image annotation improves precision and accuracy.

Data Processing

Data is processed using natural language processing (NLP). The emails are preprocessed using NLP (Natural Language

Processing) techniques and NLTK (Natural Language Tool Kit). Pre-processing email steps

a. In the emails, all uppercase letters and words are transformed to lowercase.

b. Tokenization: Using space as a separator, the email is divided into little tokens

c. Stop words are eliminated: Stop words such as and, a, and an are all removed.

d. Stemming: All verb forms tokens (V2,V3,V4) are substituted with the same token in verb form (V1).

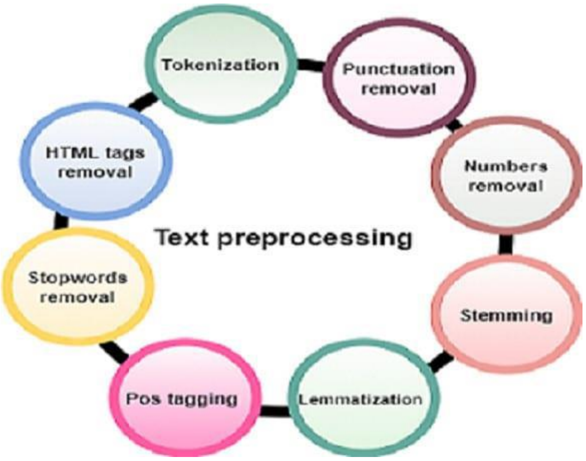


Fig :1Data processing

Sentence Embedding

The machine learning model uses a list of floats, commonly known as vectors, as input. As a result, we employed DSSM to provide input to our machine learning model (Deep Structured Semantic Model). Our preprocessed tweet was turned into a 768 dimensionVector using the Sentence Transformer Tiny Bert from Sent2Vec Vector. After the sentence is turned into a vector, the shape of the vector is (1,768).

Sentiment Analysis

The proposed system can tell if a vector is positive or negative.The MLP model is trained with 90% of training data and 10% of testing data and has a 79 percent accuracy.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n}}{\sqrt{\sum_{i=1}^{n} A_i^2}}$$

4. RESULTS AND DATA ANALYSIS

Data analysis done on the results of trained data of Support Vector machine, Long Short Term Memory and Convolution Neural Network, the accuracy we get from these 3 are compared whereas we are also training the vector data emails by the Naïve Bayes, Artificial Neural Network and Decision Tree.

Comparing all the algorithms obtained accuracy, like Support vector machine, long short term memory, convolution neural network, Naïve Bayes, Decision tree and Artificial neural networks, and also the plots are generated for the comparisons of accuracy obtained by all the above algorithms.

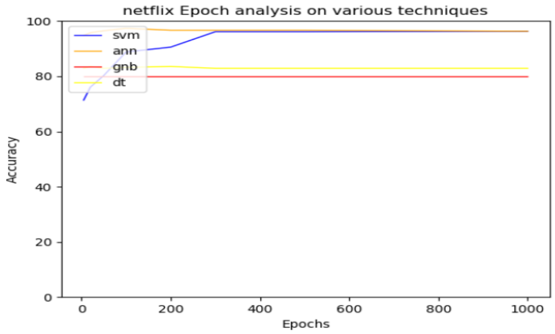
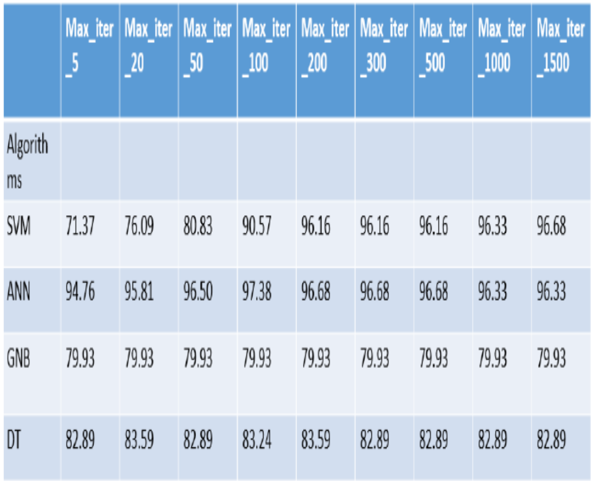


Fig:2 Comparison between 4 algorithms in graph form

This figure represents the comparison between 4 algorithms Support Vector Machine, Naïve Bayes, Artificial Neural Network and Decision Tree.



	Max_iter _5	Max_iter _20	Max_iter _50	Max_iter _100	Max_iter _200	Max_iter _300	Max_iter _500	Max_iter _1000	Max_iter _1500
Algorithms									
SVM	71.37	76.09	80.83	90.57	96.16	96.16	96.16	96.33	96.68
ANN	94.76	95.81	96.50	97.38	96.68	96.68	96.68	96.33	96.33
GNB	79.93	79.93	79.93	79.93	79.93	79.93	79.93	79.93	79.93
DT	82.89	83.59	82.89	83.24	83.59	82.89	82.89	82.89	82.89

Fig:3 Comparison between 4 algorithms in Tabular form

This is the figure which indicates comparison table of 4 algorithms Naive Bayes, Support vector machine, Decision tree and Artificial neural network. We can observe support vector machine algorithm where the epochs are increasing the accuracy is also increasing gradually, at 200 to 1000 the accuracy was constant but at 1500 epochs the accuracy as increased then after 1500 epochs the accuracy is constant.

If we observe Artificial neural network the accuracy as increased as epochs increased but at 200 and above epochs the accuracy got constant there is no farther increase in it.

If we observe Naïve Bayes algorithm as the epochs increasing there is no change in accuracy, the accuracy is constant even at low epochs and high epochs.

If we observe Decision tree algorithm as the epochs are increased the accuracy increased but with very slight changes. There is a small change with the increase of epochs.

5. CONCLUSION

Support Vector Method (SVM), Long Short-Term Memory (LSTM), and Convolution Neural Network (CNN) methods were presented for email detection. These three ML algorithms are utilized to achieve high accuracy because they are employed in classification. Header-based features, URL-based features, and script-based features have all been extracted. To train and test the data set, the Support Vector Machine, Long Short-Term Memory, and Convolution Neural Network techniques were employed. The results of the experiments reveal that our strategy outperforms others in terms of True Positive Rate and accuracy. Although our method's False Positive Rate has decreased, it is still acceptable and fair.The average of the binary properties in phishing and legal emails was also determined. The results suggest that the binary traits we collected distinguish phishing emails from authentic emails. We also used Nave Bayes, Artificial Neural Networks, and Decision True to train the email data, after which we compared the outcomes of all five algorithms and plotted the results. In general, our proposed strategy is effective at detecting phishing emails. Because we are training the data with different algorithms, the comparison is excellent, and we can also see the accuracy differences: each algorithm has a different accuracy at different epochs, implying that maximum iteration is used.Following the extraction of accuracy using various techniques We compared these accuracy values depending on different epochs, then created a graph that gives a clear clarification regarding the accuracy. Support vector machine, Naive Bayes, Decision tree, and Artificial neural network Then we draw the graph and compare the results with the accuracy values of Long Short-Term Memory and Convolution Neural Network. Finally, we look at the algorithm's greatest level of precision.

6. REFERENCES

[1] “Phishing activity trends report-second quarter 2016 [EB/OL],” bhttp://docs.apwg.org/ reports/ apwgtrendsreportq22016.pdf

[2] “APWG Phishing trends reports-second quarter 2018,” http://www.antiphishing. org/.

[3] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg and E. Almomani, “A survey of phishing
email filtering techniques”, Communications Surveys Tutorials, IEEE, vol15, iss4, pp. 2070–
2090, 2013.

[4] I. Fette, N. Sadeh and A. Tomasic, “Learning to detect phishing emails”, Proceedings of the 16th International Conference on World Wide Web, pp. 649–656, 2007.

[5] Sunil B. Rathod, Tareek M. Pattewar. Content Based Spam Detection in Email using Bayesian
Classifier [C] // International Conference on Communications and Signal Processing (ICCSP).
2015.

[6] Toolan, Fergus and Carthy, Joe, “Feature selection for spam and phishing detection,” eCrime

Researchers Summit (eCrime), 2010, pp. 1–12.

[7] XIANG G, HONG J, ROSE C P, et al. Cantina+: A feature-rich machine learning framework fordetecting phishing Web sites [J]. ACM Transactions on Information and System Security
(TISSEC), 2011, 14 (2): 21.

[8] NaghmehMoradpoor, Benjamin Clavie and Bill Buchanan. "Employing machine learning
techniques for detection and classification of phishing emails," Computing Conference, 2017.

[9] May, Lew et al. Phishing Email Detection Technique by using Hybrid Features, 9th InternationalConference on IT in Asia (CITA), 2015

[10] Saif M. Mohammad. Sentiment "Analysis of Mail and Books". Technical report, National
Research Council Canada, 2011.